

LEARNING RATES OF REGRESSION WITH q -NORM LOSS AND THRESHOLD[†]

Ting Hu

School of Mathematics and Statistics, Wuhan University
Luoja Hill, Wuhan 430072, China, tinghu@whu.edu.cn

Yuan Yao

School of Mathematical Sciences, Peking University
Beijing 100871, China, yuany@math.pku.edu.cn

Abstract

This paper studies some robust regression problems associated with the q -norm loss ($q \geq 1$) and the ϵ -insensitive q -norm loss in the reproducing kernel Hilbert space. We establish a variance-expectation bound under a priori noise condition on the conditional distribution, which is the key technique to measure the error bound. Explicit learning rates will be given under the approximation ability assumptions on the reproducing kernel Hilbert space.

Key Words and Phrases. Insensitive q -norm loss, quantile regression, reproducing kernel Hilbert space, sparsity.

Mathematical Subject Classification. 68Q32, 41A25

1 Introduction

In this paper we consider regression with the q -norm loss ψ_q with $q \geq 1$ and an ϵ -insensitive q -norm loss ψ_q^ϵ (to be defined) with a threshold $\epsilon > 0$. Here ψ_q is the univariate function defined by $\psi_q(u) = |u|^q$. For a learning algorithm generated by a regularization scheme in

reproducing kernel Hilbert spaces, learning rates and approximation error will be presented when ϵ is chosen appropriately for balancing learning rates and sparsity.

For $q = 1$, the regression problem is the classical statistical method of least absolute deviations which is more robust than the least squares method and is resistant to outliers in data [4]. Its associated loss $\psi(u) = |u|$, $u \in \mathbb{R}$, is widely used in practical applications for robustness. In fact, for all $q < 2$, the loss ψ_q is less sensitive to outliers and is thus more robust than the square loss. Vapnik [13] proposed an ϵ -insensitive loss $\psi^\epsilon(u) : \mathbb{R} \rightarrow \mathbb{R}_+$ to get sparsity in support vector regressions, which is defined by

$$\psi^\epsilon(u) = \begin{cases} |u| - \epsilon, & \text{if } |u| > \epsilon, \\ 0, & \text{if } |u| \leq \epsilon. \end{cases} \quad (1.1)$$

When fixing $\epsilon > 0$, error analysis was conducted in [12]. Xiang, Hu and Zhou [17, 18] showed how to accelerate learning rates and preserve sparsity by adapting ϵ . In [5], they discussed the convergence ability with flexible ϵ in an online algorithm. For the quantile regression with $\epsilon = 0$ and a pinball loss having different slopes in different sides of the origin in \mathbb{R} [6], Steinwart and Christmann [10, 9] established comparison theorems and derived learning rates under some noise conditions.

In this paper, we apply the q -norm loss ψ_q with $q > 1$ to improve the convexity of the insensitive loss ψ . Our results show how the insensitive parameter ϵ that produces the sparsity can be chosen adaptively as the function of the sample size $\epsilon = \epsilon(T) \rightarrow 0$ when $T \rightarrow \infty$, to affect the error rates of the learning algorithm (to be defined by (1.4)). Such results include some early studies as special cases.

In the sequel, assume that the input space X is a compact metric space and the output space $Y = \mathbb{R}$. Let ρ be a Borel probability measure on $Z := X \times Y$, $\rho_x(\cdot)$ be the conditional distribution of ρ at each $x \in X$ and ρ_X be the marginal distribution on X . For a measurable function $f : X \rightarrow Y$, the *generalization error* $\mathcal{E}(f)$ associated with the q -norm loss ψ_q , is defined by

$$\mathcal{E}(f) = \int_Z \psi_q(y - f(x)) d\rho. \quad (1.2)$$

Denote $f_q : X \rightarrow Y$ as the minimizer of the generalization error $\mathcal{E}(f)$ over all measurable functions. Its properties and the corresponding learning problem in the empirical risk minimization framework were discussed in [20]. When $q = 1$, the target function f_q is a function

containing the medians of the conditional distribution for all $x \in X$. For symmetric distributions, the median is also the regression function, which is the conditional mean for given X . We aim at learning the minimizer f_q from a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^T \in Z^T$, which is assumed to be independently drawn according to ρ . Inspired by the ϵ -insensitive loss [13], we introduce an ϵ -insensitive q -norm loss ψ_q^ϵ which is defined by

$$\psi_q^\epsilon(u) = \begin{cases} (|u| - \epsilon)^q, & \text{if } |u| > \epsilon, \\ 0, & \text{if } |u| \leq \epsilon. \end{cases} \quad (1.3)$$

Our learning task will be carried out by a regularization scheme in reproducing kernel Hilbert spaces. With a continuous, symmetric and positive semidefinite function $K : X \times X \rightarrow \mathbb{R}$ (called a Mercer kernel), the *reproducing kernel Hilbert space* (RKHS) \mathcal{H}_K is defined as the completion of the span of $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_u \rangle_K = K(x, u)$. The regularization algorithm in the paper takes the form

$$f_{\mathbf{z}}^\epsilon = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t=1}^T \psi_q^\epsilon(f(x_t) - y_t) + \lambda \|f\|_K^2 \right\}. \quad (1.4)$$

Here $\lambda > 0$ is a regularization parameter. Our learning rates are stated in terms of approximation or regularization error, noise conditions, and the capacity of the RKHS. Our main goal is to study how the learned function $f_{\mathbf{z}}^\epsilon$ in (1.4) converges to the target function f_q . There is a large literature [1, 16, 7] in learning theory for studying *the approximation error* or *regularization error* $\mathcal{D}(\lambda)$ of the triple (K, ρ, q) defined by

$$\mathcal{D}(\lambda) = \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_q) + \lambda \|f\|_K^2 \right\}, \quad \lambda > 0.$$

The regularization function is defined as

$$f_\lambda = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_q) + \lambda \|f\|_K^2 \right\}. \quad (1.5)$$

In the sequel, let $L_{\rho_X}^p$ with $p > 0$ be the space of p integrable functions with respect to ρ_X and $\|\cdot\|_{L_{\rho_X}^p}$ be the norm in $L_{\rho_X}^p$. A usual assumption on the regularization error $\mathcal{D}(\lambda)$ which imposes certain smoothness on \mathcal{H}_K is

$$\mathcal{D}(\lambda) \leq \mathcal{D}_0 \lambda^\beta, \quad \forall \lambda > 0 \quad (1.6)$$

with some $0 < \beta \leq 1$ and $\mathcal{D}_0 > 0$.

Remark 1. Assumption (1.6) always holds with $\beta = 0$. When the target function $f_q \in \mathcal{H}_K$ and \mathcal{H}_K is dense in $C(X)$ which consists of bounded continuous functions on X , the approximation error $\mathcal{D}(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$. Thus, the decay (1.6) is natural and can be illustrated in terms of interpolation spaces [7]. Define the integral operator $L_K : L_{\rho_X}^2 \rightarrow L_{\rho_X}^2$ by $L_K(f)(x) = \int_X K(x, y)f(y)d\rho_X, x \in X, f \in L_{\rho_X}^2$ and suppose that the minimizer f_q is in the range of L_K^ν with $0 < \nu \leq \frac{1}{2}$. When $q = 1$, the approximation error $\mathcal{D}(\lambda)$ can be $O(\lambda^{\frac{\nu}{1-\nu}})$ for quantile regression [18]. When $q = 2$, $\mathcal{D}(\lambda) = O(\lambda^{2\nu})$ for the least square. For other $q > 1$, the associated loss ψ_q is Lipschitz in a bounded domain and the corresponding $\mathcal{D}(\lambda)$ can be characterized by the \mathcal{K} -functional [1], which can have the same polynomial decay as (1.6).

We assume that the conditional distribution $\rho_x(\cdot)$ is supported on $[-M, M]$, $M > 0$ at each x and is non-degenerate, i.e. any non-empty open set of Y has strictly positive measure, which ensures that the target function f_q is unique. Without loss of generality, let the support of $\rho_x(\cdot)$ be $[-\frac{1}{2}, \frac{1}{2}]$ at each $x \in X$ and our analysis below is applicable for any $M > 0$. We will prove that in the next section. It is natural to project values of the learned function $f_{\mathbf{z}}^\epsilon$ onto some interval by the projection operator [1, 15].

Definition 1. The projection operator π on the space of measurable functions $f : X \rightarrow \mathbb{R}$ onto the interval $[-1, 1]$ is defined by

$$\pi(f(x)) = \begin{cases} 1, & \text{if } f(x) \geq 1, \\ f(x), & \text{if } -1 < f(x) < 1, \\ -1, & \text{if } f(x) \leq -1. \end{cases}$$

To demonstrate our main result in the general case, we shall give the following learning rate in the special case when K is C^∞ .

Theorem 1. Let $X \subset \mathbb{R}^n$ and $K \in C^\infty(X \times X)$. Assume that $f_q \in \mathcal{H}_K$ with $q > 1$, $\|f_q\|_\infty \leq \frac{1}{4}$ and the conditional distributions $\{\rho_x(\cdot)\}_{x \in X}$ have density functions given by

$$\frac{d\rho_x}{dy}(y) = \begin{cases} A|y - f_q(x)|^\varphi, & \text{if } |y - f_q(x)| \leq \frac{1}{4}, \\ 0, & \text{otherwise,} \end{cases} \quad (1.7)$$

where $A = 2^{2\varphi+1}(\varphi + 1)$, $\varphi > 0$. Take $\lambda = \epsilon = T^{-\frac{q+\varphi+1}{2(q+\varphi)}}$, then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\|\pi(f_{\mathbf{z}}^\epsilon) - f_q\|_{L_{\rho_X}^{q+\varphi+1}} \leq C' \sqrt{\log \frac{3}{\delta}} T^{-\frac{1}{2(q+\varphi)}},$$

where C' is a constant independent of T or δ .

To state our main result in the general case, we need a noise condition on the measure ρ introduced in [9, 10].

Definition 2. Let $0 < p \leq \infty$ and $w > 0$. We say that ρ has a p -average type w if there exist two functions b and a from X to \mathbb{R} such that $\{ba^w\}^{-1} \in L_{\rho_X}^p$ and for any $x \in X$ and $s \in (0, a(x)]$, there holds

$$\rho_x(\{y : f_q(x) \leq y \leq f_q(x) + s\}) \geq b(x)s^w$$

and

$$\rho_x(\{y : f_q(x) - s \leq y \leq f_q(x)\}) \geq b(x)s^w. \quad (1.8)$$

This assumption can be satisfied by many common conditional distributions such as Guassians, students' t distributions and uniform distributions. In the following, we will give an example to illustrate Definition 2 in detail. More examples can be found in [9, 10].

Example 1. We assume that the conditional distributions $\{\rho_x(\cdot)\}_{x \in X}$ are Gaussian distributions with a uniform variance $\sigma > 0$, i.e. $\frac{d\rho_x}{dy}(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y-u_x)^2}{2\sigma^2}\}$, where $\{u_x\}_{x \in X}$ are expectations of the Gaussian distributions $\{\rho_x(\cdot)\}_{x \in X}$. It is not difficult to check that the minimizer $f_\rho(x)$ can take the value of u_x at each $x \in X$, then for any $s \in (0, \sigma]$, there holds

$$\begin{aligned} \rho_x(\{y : f_q(x) \leq y \leq f_q(x) + s\}) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{f_\rho(x)}^{f_\rho(x)+s} \exp\{-\frac{(y-u_x)^2}{2\sigma^2}\} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_0^s \exp\{-\frac{y^2}{2\sigma^2}\} dy \geq \frac{1}{\sqrt{2\pi}\sigma} \int_0^s \exp\{-\frac{s^2}{2\sigma^2}\} dy \geq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\sigma} s. \end{aligned}$$

By similarity, we also have that $\rho_x(\{y : f_q(x) - s \leq y \leq f_q(x)\}) \geq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\sigma} s$. Thus, the measure ρ has a ∞ -average type 1.

Our error analysis is related to the capacity of the hypothesis space \mathcal{H}_K which is measured by covering numbers.

Definition 3. For a subset S of $C(X)$ and $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon)$ is the minimal integer $l \in \mathbb{N}$ such that there exist l disks with radius ε covering S .

The covering numbers of balls $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$ with $R > 0$ of the RKHS have been well understood in the learning theory [22, 23]. In this paper, we assume for some $k > 0$ and $C_k > 0$ that

$$\log \mathcal{N}(B_1, \varepsilon) \leq C_k \left(\frac{1}{\varepsilon}\right)^k, \quad \forall \varepsilon > 0. \quad (1.9)$$

Remark 2. When X is a bounded subset of \mathbb{R}^n and the RKHS \mathcal{H}_K is a Sobolev space $H^m(X)$ with index m , it is shown [22] that the condition (1.9) holds true with $k = \frac{2n}{m}$. If the kernel K lies in the smooth space $C^\infty(X \times X)$, then (1.9) is satisfied for an arbitrarily small $k > 0$. Another common way to measure the capacity of \mathcal{H}_K is the empirical covering number [21], which is out of scope of our discussion in this paper.

Denote

$$\theta = \min\left\{\frac{2}{q+w}, \frac{p}{p+1}\right\} \in (0, 1], \quad r = \frac{p(q+w)}{p+1} > 0. \quad (1.10)$$

The following learning rates in the general case will be proved in Section 4. One need to point out that the proof of Theorem 2 is only applicable to the case $q > 1$. However, when $q = 1$, it is a special case of quantile regression and the same learning rates as those of Theorem 2 can be found in [17, 18].

Theorem 2. Suppose that ρ has a p -average type w for some $0 < p \leq \infty$ and $\omega > 0$. Assume that the regularization error condition (1.6) is satisfied for some $0 < \beta \leq 1$ and (1.9) holds with $k > 0$. Take $\lambda = T^{-\alpha}$, $\epsilon = T^{-\eta}$ with $0 < \alpha \leq 1$, $0 < \eta \leq \infty$. Let $\xi > 0$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|\pi(f_{\mathbf{z}}^{\epsilon}) - f_q\|_{L_{\rho_X}^r} \leq C^* \left(\log \frac{3}{\xi}\right)^2 \sqrt{\log \frac{3}{\delta}} T^{-\Lambda} \quad (1.11)$$

where C^* is a constant independent of T or δ ,

$$\Lambda = \frac{1}{q+w} \min \left\{ \eta, \alpha\beta, 1 - \frac{q(1-\beta)\alpha}{2}, \frac{1}{2-\theta}, \frac{1}{2+k-\theta} - \frac{k}{1+k} \vartheta \right\}$$

with $\vartheta = \max \left\{ \frac{\alpha-\eta}{2}, \frac{\alpha(1-\beta)}{2}, \frac{\alpha}{2} + \frac{q(1-\beta)\alpha}{4} - \frac{1}{2}, \frac{\alpha}{2} - \frac{1}{2(2-\theta)}, \frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)} + \xi \right\} \geq 0$ provided that

$$\vartheta < \frac{1+k}{k(2+k-\theta)}. \quad (1.12)$$

Corollary 1. Let $X \subset \mathbb{R}^n$, $K \in C^\infty(X \times X)$. Assume (1.6) and (1.8). Take $\lambda = T^{-1}$, $\epsilon = T^{-\eta}$ with $0 < \eta \leq \infty$. If $1 < q \leq 2$, then the index Λ for the learning rate (1.11) is $\frac{1}{q+w} \min\{\eta, \beta, \frac{1}{2-\theta}\}$.

Remark 3. When $\eta = \infty$, the corresponding threshold ϵ is 0 and it is a least square problem for $q = 2$, which is widely discussed in [15, 16]. If ρ has a ∞ -average type w with $w > 0$ and $f_q \in \mathcal{H}_K$, the learning rate $\|\pi(f_{\mathbf{z}}^\epsilon) - f_q\|_{L_{\rho_X}^{2+w}} = O(T^{-\frac{1}{2(1+w)}})$ for the least square. It follows that the error $\|\pi(f_{\mathbf{z}}^\epsilon) - f_q\|_{L_{\rho_X}^2}^2 = O(T^{-\frac{1}{1+w}})$ by $\|\cdot\|_{L_{\rho_X}^2} \leq \|\cdot\|_{L_{\rho_X}^{2+w}}$. Thus, it can be near the optimal rate $O(T^{-1})$ in $L_{\rho_X}^2$ space if w is small enough.

When $1 < q < 2$, the learning error will be $O(T^{-\frac{1}{q+w} \min\{\beta, \frac{q+w}{2(q+w-1)}, \frac{p+1}{p+2}\}})$ with choice $\eta \geq \beta$, depending only on the \mathcal{H}_K 's approximation ability (1.6) and noise condition (1.8). Specially, when q goes to 1, it is the quantile regression [17, 18] and the best rate is $O(T^{-\frac{1}{1+w}})$ in this paper if ρ has a ∞ -average type w with $0 < w \leq 1$ and $f_q \in \mathcal{H}_K$.

2 Comparison and Perturbation Theorem

Approximation or learning ability of a regularized algorithm for regression problems can usually be studied by estimating the *excess generalization error* $\mathcal{E}(f) - \mathcal{E}(f_q)$ for the learned function $f_{\mathbf{z}}^\epsilon$ from the algorithm (1.4). However the following comparison theorem would yield bounds for the error $\|f - f_q\|_{L_{\rho_X}^r}$ in the space $L_{\rho_X}^r$ when the noise condition is satisfied.

Theorem 3. If ρ has a p -average type w , then for any measurable function $f : X \rightarrow [-1, 1]$ we have the inequality

$$\|f - f_q\|_{L_{\rho_X}^r} \leq C_r (\mathcal{E}(f) - \mathcal{E}(f_q))^{\frac{1}{q+w}} \quad (2.1)$$

where the constant $C_r = 2^{\frac{q-1}{q+w}} q^{-\frac{1}{q+w}} (q+w)^{\frac{1}{q+w}} \|(ba^w)^{-1}\|_{L_{\rho_X}^p}^{\frac{1}{q+w}}$.

Proof. For a measurable function $f : X \rightarrow [-1, 1]$, the generalization error $\mathcal{E}(f)$ is rewritten as $\mathcal{E}(f) = \int_X C_{q,x}(f(x)) d\rho_X$ where

$$C_{q,x}(t) = \int_Y \psi_q(y-t) d\rho_x(y) = \int_{y>t} (y-t)^q d\rho_x(y) + \int_{y<t} (t-y)^q d\rho_x(y), \quad x \in X.$$

Denote $t_x^* = \min_{t \in \mathbb{R}} C_{q,x}(t)$. It is obvious that the minimizer $f_q(x)$ of $\mathcal{E}(f)$ takes the value of t_x^* for each $x \in X$. Noting that the conditional distribution $\rho_x(\cdot)$ is supported on $[-\frac{1}{2}, \frac{1}{2}]$, the minimizer t_x^* can be on $[-\frac{1}{2}, \frac{1}{2}]$. Consider the case $q > 1$. Since the loss function ψ_q is differential and $|\frac{d\psi_q(y-t)}{dt}| \leq q|y-t|^{q-1} \leq q$ for all $y, t \in [-\frac{1}{2}, \frac{1}{2}]$, by the corollary of Lebesgue control convergence theorem, we can exchange the order of of integration and derivation of $C'_{q,x}(t)$ as $C'_{q,x}(t) = \frac{d}{dt} \int_Y \psi_q(y-t) d\rho_x(y) = \int_Y \frac{d\psi_q(y-t)}{dt} d\rho_x(y)$. This together with the fact $C'_{q,x}(t_x^*) = 0, \forall x \in X$, we have

$$C'_{q,x}(t_x^*) = q \int_{y<t_x^*} (t_x^* - y)^{q-1} d\rho_x(y) - q \int_{y>t_x^*} (y - t_x^*)^{q-1} d\rho_x(y) = 0,$$

which means that

$$\int_{y<t_x^*} (t_x^* - y)^{q-1} d\rho_x(y) = \int_{y>t_x^*} (y - t_x^*)^{q-1} d\rho_x(y) \quad (2.2)$$

Let $t_x^* = 0$ for simply, then we have $C_{q,x}(t) - C_{q,x}(0) = \int_0^t C'_{q,x}(s) ds, \forall t > 0$. Noting that for $s > 0$,

$$\begin{aligned} C'_{q,x}(s) &= q \left(\int_{y<s} (s-y)^{q-1} d\rho_x(y) - \int_{y>s} (y-s)^{q-1} d\rho_x(y) \right) \\ &= q \left(\int_{y<0} (s-y)^{q-1} d\rho_x(y) + \int_{0 \leq y < s} (s-y)^{q-1} d\rho_x(y) - \int_{y>s} (y-s)^{q-1} d\rho_x(y) \right) \\ &\geq q \left(\int_{y<0} (-y)^{q-1} d\rho_x(y) + \int_{0 \leq y < s} (s-y)^{q-1} d\rho_x(y) - \int_{y>s} (y-s)^{q-1} d\rho_x(y) \right). \end{aligned}$$

The above first term together with (2.2), then

$$\begin{aligned}
C'_{q,x}(s) &\geq q \left(\int_{y>0} y^{q-1} d\rho_x(y) + \int_{0 \leq y < s} (s-y)^{q-1} d\rho_x(y) - \int_{y>s} (y-s)^{q-1} d\rho_x(y) \right) \\
&\geq q \left(\int_{y>0} y^{q-1} d\rho_x(y) + \int_{0 \leq y < s} (s-y)^{q-1} d\rho_x(y) - \int_{y>s} y^{q-1} d\rho_x(y) \right) \\
&= q \left(\int_{0 < y \leq s} y^{q-1} d\rho_x(y) + \int_{0 \leq y < s} (s-y)^{q-1} d\rho_x(y) \right) \\
&= q \int_{0 \leq y \leq s} \left(y^{q-1} + (s-y)^{q-1} \right) d\rho_x(y) \geq 2^{1-q} q s^{q-1} \rho_x(\{y : 0 \leq y \leq s\}).
\end{aligned}$$

Thus,

$$C_{q,x}(t) - C_{q,x}(0) \geq 2^{1-q} \cdot q \int_0^t s^{q-1} \rho_x(\{y : 0 \leq y \leq s\}) ds.$$

Let us consider the first case $t \in [0, a(x)]$. Noting the noise condition (1.8) and $a(x) \leq 1$, we obtain that

$$C_{q,x}(t) - C_{q,x}(0) \geq 2^{1-q} \cdot q \int_0^t s^{q-1} b(x) s^w ds = \frac{2^{1-q} q}{q+w} b(x) t^{q+w} \geq \frac{2^{1-q} q}{q+w} b(x) a(x)^w t^{q+w}.$$

For the second case $t \in [a(x), 1]$, we have

$$\begin{aligned}
C_{q,x}(t) - C_{q,x}(0) &\geq 2^{1-q} \cdot q \left(\int_0^{a(x)} s^{q-1} \rho_x(\{y : 0 \leq y \leq s\}) ds + \int_{a(x)}^t s^{q-1} \rho_x(\{y : 0 \leq y \leq s\}) ds \right) \\
&\geq 2^{1-q} \cdot q \left(\int_0^{a(x)} s^{q-1} b(x) s^w ds + \int_{a(x)}^t s^{q-1} b(x) a(x)^w ds \right) \\
&= 2^{1-q} \cdot q \left(\frac{b(x) a(x)^w t^q}{q} - \frac{w b(x) a(x)^{q+w}}{q(q+w)} \right) \geq 2^{1-q} \left(b(x) a(x)^w t^q - \frac{w b(x) a(x)^w t^q}{q+w} \right) \\
&= \frac{2^{1-q} q}{q+w} b(x) a(x)^w t^q \geq \frac{2^{1-q} q}{q+w} b(x) a(x)^w t^{q+w}.
\end{aligned}$$

In general, we can see that for any $0 < t \leq 1$,

$$C_{q,x}(t) - C_{q,x}(0) \geq \frac{2^{1-q} q}{q+w} b(x) a(x)^w t^{q+w}. \quad (2.3)$$

By similarity, if $-1 \leq t < 0$, we also have

$$C_{q,x}(t) - C_{q,x}(0) \geq \frac{2^{1-q} q}{q+w} b(x) a(x)^w t^{q+w}. \quad (2.4)$$

Applying the two above inequalities (2.3) and (2.4) with $t = f(x)$ and $t_x^* = f_q(x)$, we have that

$$|f(x) - f_q(x)|^{q+w} \leq 2^{q-1} q^{-1} (q+w) (b(x)a(x)^w)^{-1} (C_{q,x}(f(x)) - C_{q,x}(f_q(x))).$$

By $\frac{p}{p+1}$ power and integration,

$$\begin{aligned} \int_X |f(x) - f_q(x)|^{\frac{p(q+w)}{p+1}} d\rho_X &\leq 2^{\frac{p(q-1)}{p+1}} q^{-\frac{p}{p+1}} (q+w)^{\frac{p}{p+1}} \\ &\int_X [(b(x)a(x)^w)^{-1}]^{\frac{p}{p+1}} (C_{q,x}(f(x)) - C_{q,x}(f_q(x)))^{\frac{p}{p+1}} d\rho_X. \end{aligned}$$

This with Holder inequality $\|\cdot\|_{L_{\rho_X}^1} \leq \|\cdot\|_{L_{\rho_X}^{p^*}} \|\cdot\|_{L_{\rho_X}^{q^*}}$, $\frac{1}{p^*} + \frac{1}{q^*} = 1$, we obtain that for $p^* = p+1$ and $q^* = \frac{p+1}{p}$,

$$\int_X |f(x) - f_q(x)|^{\frac{p(q+w)}{p+1}} d\rho_X \leq 2^{\frac{p(q-1)}{p+1}} q^{-\frac{p}{p+1}} (q+w)^{\frac{p}{p+1}} \|(ba^w)^{-1}\|_{L_{\rho_X}^p}^{\frac{p}{p+1}} (\mathcal{E}(f) - \mathcal{E}(f_q))^{\frac{p}{p+1}}.$$

Then the desired conclusion (2.1) holds. For $q = 1$, (2.1) also holds and the proof can be found in [18]. \square

It yields a variance-expectation bound which will be applied in the next section.

Lemma 1. *Under the same conditions as Theorem 3, for any measurable function $f : X \rightarrow [-1, 1]$, we have the inequality*

$$\mathbb{E}\{(\psi_q(f(x) - y) - \psi_q(f_q(x) - y))^2\} \leq C_\theta (\mathcal{E}(f) - \mathcal{E}(f_q))^\theta \quad (2.5)$$

where the power index θ is defined as (1.10) and $C_\theta = C_r^2 + 2^{2-r}(1 + \|f_q\|_\infty^{2-r})C_r^r$.

Proof. By the continuity of $\psi_q(u)$ and $|y| \leq \frac{1}{2}$, we see that

$$\begin{aligned} |\psi_q(f(x) - y) - \psi_q(f_q(x) - y)| &\leq q(\|f\|_\infty^{q-1} + \|f_q\|_\infty^{q-1} + 1)|f(x) - f_q(x)| \\ &\leq q(2 + \|f_q\|_\infty^{q-1})|f(x) - f_q(x)|. \end{aligned}$$

It implies that

$$\mathbb{E}\{(\psi_q(f(x) - y) - \psi_q(f_q(x) - y))^2\} \leq q^2(\|f_q\|_\infty^{q-1} + 2)^2 \mathbb{E}|f(x) - f_q(x)|^2.$$

If $r > 2$, then

$$\mathbb{E}|f(x) - f_q(x)|^2 \leq \{\mathbb{E}|f(x) - f_q(x)|^r\}^{\frac{2}{r}} \leq C_r^2(\mathcal{E}(f) - \mathcal{E}(f_q))^{\frac{2}{q+w}}.$$

Else,

$$\begin{aligned} \mathbb{E}|f(x) - f_q(x)|^2 &\leq \mathbb{E}\{|f(x) - f_q(x)|^{2-r} \cdot |f(x) - f_q(x)|^r\} \\ &\leq 2^{2-r}(\|f\|_\infty^{2-r} + \|f_q\|_\infty^{2-r})\mathbb{E}|f(x) - f_q(x)|^r \\ &\leq 2^{2-r}(1 + \|f_q\|_\infty^{2-r})C_r^r(\mathcal{E}(f) - \mathcal{E}(f_q))^{\frac{p}{p+1}}. \end{aligned}$$

Combining the above two cases, we can get the conclusion (2.5). \square

The threshold ϵ changes with the sample size $n = n(T)$ and plays a crucial role in the design of algorithm (1.4). By Taylor expansion, we have the following relation

$$\psi_q^\epsilon(u) \leq \psi(u) \leq \psi_q^\epsilon(u) + q|u|^{q-1}\epsilon, \quad \forall u \in \mathbb{R}. \quad (2.6)$$

When the threshold $\epsilon \rightarrow 0$, the ϵ -insensitive q -norm loss ψ_q^ϵ converges to the q -norm function ψ_q almost surely. In the following, we shall study the approximation of the target function f_q by f_q^ϵ which is the minimizer of the ϵ -generalization error $\mathcal{E}^\epsilon(f) = \int_Z \psi_q^\epsilon(f(x) - y)d\rho$ for $\epsilon > 0$. Denote

$$C_{q,x}^\epsilon(t) = \int_Y \psi_q^\epsilon(y-t)d\rho_x(y) = \int_{y>t+\epsilon} (y-t-\epsilon)^q d\rho_x(y) + \int_{y<t-\epsilon} (t-y-\epsilon)^q d\rho_x(y), \quad x \in X. \quad (2.7)$$

and t_x^ϵ is the minimizer of $C_{q,x}^\epsilon(t)$. By the same proof procedure as (2.2) in Theorem 3, we also get

$$\int_{y>t_x^\epsilon+\epsilon} (y - t_x^\epsilon - \epsilon)^{q-1} d\rho_x(y) = \int_{y<t_x^\epsilon-\epsilon} (t_x^\epsilon - \epsilon - y)^{q-1} d\rho_x(y) \quad (2.8)$$

and f_q^ϵ takes the value of t_x^ϵ at each $x \in X$. Then the perturbation properties hold. We use some ideas from [3] in the proof.

Proposition 1. For $\epsilon > 0$, then

$$\|f_q^\epsilon - f_q\|_\infty \leq \epsilon. \quad (2.9)$$

For any measurable function f on X , we have

$$\mathcal{E}(f) - \mathcal{E}(f_q) \leq \mathcal{E}^\epsilon(f) - \mathcal{E}^\epsilon(f_q^\epsilon) + q(\|f\|_\infty^{q-1} + 1)\epsilon \quad (2.10)$$

Proof. Suppose that there exist a $x \in X$ satisfying $f_q^\epsilon(x) - f_q(x) > \epsilon$. Consider the case $q > 1$. Together with the fact (2.2) and $t_x^* = f_q(x)$, we note that

$$\begin{aligned} \int_{y < f_q^\epsilon(x) - \epsilon} (f_q^\epsilon(x) - \epsilon - y)^{q-1} d\rho_x(y) &\geq \int_{y < f_q(x)} (f_q^\epsilon(x) - \epsilon - y)^{q-1} d\rho_x(y) \\ &\geq \int_{y < f_q(x)} (f_q(x) - y)^{q-1} d\rho_x(y) = \int_{y > f_q(x)} (y - f_q(x))^{q-1} d\rho_x(y) \end{aligned} \quad (2.11)$$

It is obvious that $f_q^\epsilon(x) + \epsilon > f_q(x)$ by the hypothesis that $f_q^\epsilon(x) - f_q(x) > \epsilon$ for any $\epsilon > 0$. By (2.8) with $t_x^\epsilon = f_q^\epsilon(x)$, we also get

$$\begin{aligned} \int_{y > f_q(x)} (y - f_q(x))^{q-1} d\rho_x(y) &\geq \int_{y > f_q^\epsilon(x) + \epsilon} (y - f_q(x))^{q-1} d\rho_x(y) \\ &\geq \int_{y > f_q^\epsilon(x) + \epsilon} (y - f_q^\epsilon(x) - \epsilon)^{q-1} d\rho_x(y) = \int_{y < f_q^\epsilon(x) - \epsilon} (f_q^\epsilon(x) - \epsilon - y)^{q-1} d\rho_x(y). \end{aligned} \quad (2.12)$$

Combining (2.12) with (2.11), we know that

$$\begin{aligned} \int_{y < f_q^\epsilon(x) - \epsilon} (f_q^\epsilon(x) - \epsilon - y)^{q-1} d\rho_x(y) &= \int_{y < f_q(x)} (f_q^\epsilon(x) - \epsilon - y)^{q-1} d\rho_x(y) \\ &= \int_{y < f_q(x)} (f_q(x) - y)^{q-1} d\rho_x(y) = \int_{y > f_q(x)} (y - f_q(x))^{q-1} d\rho_x(y) \\ &= \int_{y > f_q^\epsilon(x) + \epsilon} (y - f_q(x))^{q-1} d\rho_x(y) = \int_{y > f_q^\epsilon(x) + \epsilon} (y - f_q^\epsilon(x) - \epsilon)^{q-1} d\rho_x(y) \end{aligned} \quad (2.13)$$

The above equalities hold if and only if $\rho_x(\{y : y > f_q(x)\}) = 0$ and $\rho_x(\{y : y < f_q^\epsilon(x) - \epsilon\}) = 0$ at the same time. Immediately, we see that $\rho_x(\{y : y \leq f_q(x)\}) = 1 - \rho_x(\{y : y > f_q(x)\}) = 1$. By the hypothesis $f_q^\epsilon(x) - f_q(x) > \epsilon$, it follows that

$$\rho_x(\{y : y \leq f_q(x)\}) \leq \rho_x(\{y : y < f_q^\epsilon(x) - \epsilon\}) = 0.$$

This is contradiction. By similarity, we get that $f_q^\epsilon(x) - f_q(x) < -\epsilon$ for each $x \in X$. Then the desired conclusion (2.9) holds. By the relation (2.6) and $|y| \leq \frac{1}{2}$, we can see that

$$\mathcal{E}(f) - \mathcal{E}(f_q) \leq \mathcal{E}^\epsilon(f) - \mathcal{E}^\epsilon(f_q) + q\|f - y\|_\infty^{q-1}\epsilon \leq \mathcal{E}^\epsilon(f) - \mathcal{E}^\epsilon(f_q) + q(\|f\|_\infty^{q-1} + 1)\epsilon.$$

Then the desired conclusion (2.10) holds. \square

We recall the fact that the conditional distribution $\rho_x(\cdot)$ is non-degenerate for each $x \in X$, then the uniqueness of the minimizer f_q^ϵ is stated as following. For simply, we denote f_q^ϵ as the target function f_q and $\mathcal{E}^\epsilon(f)$ as the generalization error $\mathcal{E}(f)$ with the q -norm loss ψ_q when $\epsilon = 0$ in the next proposition.

Proposition 2. *For $0 \leq \epsilon \leq \frac{1}{2}$, the function f_q^ϵ is the unique minimizer of the ϵ -generalization error $\mathcal{E}^\epsilon(f)$.*

Proof. Suppose that f_q^ϵ is not the unique minimizer. For some $x \in X$, there exists $t_1(x) < t_2(x)$ such that they are both the minimizers of $C_{q,x}^\epsilon(t)$ by (2.7) and satisfy the equality (2.8) with $t_x^\epsilon = t_1(x)$ or $t_x^\epsilon = t_2(x)$. Applying (2.8) with $t_x^\epsilon = t_1(x)$ and $t_1(x) < t_2(x)$, it follows that

$$\begin{aligned} \int_{y < t_2(x) - \epsilon} (t_2(x) - \epsilon - y)^{q-1} d\rho_x(y) &\geq \int_{y < t_1(x) - \epsilon} (t_2(x) - \epsilon - y)^{q-1} d\rho_x(y) \\ &\geq \int_{y < t_1(x) - \epsilon} (t_1(x) - \epsilon - y)^{q-1} d\rho_x(y) = \int_{y > t_1(x) + \epsilon} (y - t_1(x) - \epsilon)^{q-1} d\rho_x(y) \\ &\geq \int_{y > t_2(x) + \epsilon} (y - t_1(x) - \epsilon)^{q-1} d\rho_x(y) \geq \int_{y > t_2(x) + \epsilon} (y - t_2(x) - \epsilon)^{q-1} d\rho_x(y). \end{aligned}$$

Applying (2.8) with $t_x^\epsilon = t_2(x)$ again, we see that the first term of the above inequality $\int_{y < t_2(x) - \epsilon} (t_2(x) - \epsilon - y)^{q-1} d\rho_x(y)$ is equal to the last term $\int_{y > t_2(x) + \epsilon} (y - t_2(x) - \epsilon)^{q-1} d\rho_x(y)$. This implies

$$\begin{aligned} \int_{y < t_2(x) - \epsilon} (t_2(x) - \epsilon - y)^{q-1} d\rho_x(y) &= \int_{y < t_1(x) - \epsilon} (t_2(x) - \epsilon - y)^{q-1} d\rho_x(y) \\ &= \int_{y < t_1(x) - \epsilon} (t_1(x) - \epsilon - y)^{q-1} d\rho_x(y) = \int_{y > t_1(x) + \epsilon} (y - t_1(x) - \epsilon)^{q-1} d\rho_x(y) \\ &= \int_{y > t_2(x) + \epsilon} (y - t_1(x) - \epsilon)^{q-1} d\rho_x(y) = \int_{y > t_2(x) + \epsilon} (y - t_2(x) - \epsilon)^{q-1} d\rho_x(y). \end{aligned}$$

The above equalities hold if and only if $\rho_x(\{y : y < t_2(x) - \epsilon\}) = 0$ and $\rho_x(\{y : y > t_1(x) + \epsilon\}) = 0$ simultaneously. Since $\rho_x(\cdot)$ is non-degenerate and supported on $[-\frac{1}{2}, \frac{1}{2}]$, then the values of $t_1(x)$ and $t_2(x)$ must satisfy $t_2(x) - \epsilon \leq -\frac{1}{2}$ and $t_1(x) + \epsilon \geq \frac{1}{2}$. By the hypothesis $t_1(x) < t_2(x)$, we get $\epsilon > \frac{1}{2}$. This is contradict with $0 \leq \epsilon \leq \frac{1}{2}$. The proof is completed. \square

3 Error Decomposition and Sample Error

Now we can conduct an error decomposition.

Lemma 2. Define f_λ by (1.5). Let $0 \leq \epsilon \leq \frac{1}{2}$, then

$$\mathcal{E}(\pi(f_\mathbf{z}^\epsilon)) - \mathcal{E}(f_q) + \lambda \|f_\mathbf{z}^\epsilon\|_K^2 \leq S_1 + S_2 + \mathcal{D}(\lambda) + q2^{q-1}\epsilon, \quad (3.1)$$

where

$$S_1 = [\mathcal{E}(\pi(f_\mathbf{z}^\epsilon)) - \mathcal{E}(f_q)] - [\mathcal{E}_\mathbf{z}(\pi(f_\mathbf{z}^\epsilon)) - \mathcal{E}_\mathbf{z}(f_q)], \quad (3.2)$$

$$S_2 = [\mathcal{E}_\mathbf{z}(f_\lambda) - \mathcal{E}_\mathbf{z}(f_q)] - [\mathcal{E}(f_\lambda) - \mathcal{E}(f_q)]. \quad (3.3)$$

Proof. By the same procedure in [11, 14, 15, 16], $\mathcal{E}(\pi(f_\mathbf{z}^\epsilon)) - \mathcal{E}(f_q) + \lambda \|f_\mathbf{z}^\epsilon\|_K^2$ can be expressed as

$$\begin{aligned} & \{\mathcal{E}(\pi(f_\mathbf{z}^\epsilon)) - \mathcal{E}_\mathbf{z}(\pi(f_\mathbf{z}^\epsilon))\} + \{[\mathcal{E}_\mathbf{z}(\pi(f_\mathbf{z}^\epsilon)) + \lambda \|f_\mathbf{z}^\epsilon\|_K^2] - [\mathcal{E}_\mathbf{z}(f_\lambda) + \lambda \|f_\lambda\|_K^2]\} \\ & + \{\mathcal{E}_\mathbf{z}(f_\lambda) - \mathcal{E}(f_\lambda)\} + \{\mathcal{E}(f_\lambda) - \mathcal{E}(f_q) + \lambda \|f_\lambda\|_K^2\}. \end{aligned}$$

The relation (2.6) yields

$$\begin{aligned} \mathcal{E}_\mathbf{z}(\pi(f_\mathbf{z}^\epsilon)) &= \frac{1}{T} \sum_{i=1}^T \psi_q(\pi(f_\mathbf{z}^\epsilon)(x_i) - y_i) \\ &\leq \frac{1}{T} \sum_{i=1}^T \psi_q^\epsilon(\pi(f_\mathbf{z}^\epsilon)(x_i) - y_i) + q(\|\pi(f_\mathbf{z}^\epsilon)\|_\infty + |y|)^{q-1}\epsilon \\ &\leq \mathcal{E}_\mathbf{z}^\epsilon(\pi(f_\mathbf{z}^\epsilon)) + q2^{q-1}\epsilon \end{aligned} \quad (3.4)$$

and

$$\mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\lambda}) = \frac{1}{T} \sum_{i=1}^T \psi_q^{\epsilon}(f_{\lambda}(x_i) - y_i) \leq \frac{1}{T} \sum_{i=1}^T \psi_q(f_{\lambda}(x_i) - y_i) = \mathcal{E}_{\mathbf{z}}(f_{\lambda}). \quad (3.5)$$

The restriction $0 \leq \epsilon \leq \frac{1}{2}$ implies $\mathcal{E}_{\mathbf{z}}^{\epsilon}(\pi(f_{\mathbf{z}}^{\epsilon})) \leq \mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\mathbf{z}}^{\epsilon})$. By (3.4) and (3.5), then we have

$$\begin{aligned} & [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}^{\epsilon})) + \lambda \|f_{\mathbf{z}}^{\epsilon}\|_K^2] - [\mathcal{E}_{\mathbf{z}}(f_{\lambda}) + \lambda \|f_{\lambda}\|_K^2] \\ & \leq [\mathcal{E}_{\mathbf{z}}^{\epsilon}(\pi(f_{\mathbf{z}}^{\epsilon})) + \lambda \|f_{\mathbf{z}}^{\epsilon}\|_K^2] - [\mathcal{E}_{\mathbf{z}}(f_{\lambda}) + \lambda \|f_{\lambda}\|_K^2] + q2^{q-1}\epsilon \\ & \leq [\mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\mathbf{z}}^{\epsilon}) + \lambda \|f_{\mathbf{z}}^{\epsilon}\|_K^2] - [\mathcal{E}_{\mathbf{z}}(f_{\lambda}) + \lambda \|f_{\lambda}\|_K^2] + q2^{q-1}\epsilon \\ & \leq [\mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\mathbf{z}}^{\epsilon}) + \lambda \|f_{\mathbf{z}}^{\epsilon}\|_K^2] - [\mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\lambda}) + \lambda \|f_{\lambda}\|_K^2] + q2^{q-1}\epsilon. \end{aligned}$$

Since $[\mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\mathbf{z}}^{\epsilon}) + \lambda \|f_{\mathbf{z}}^{\epsilon}\|_K^2] - [\mathcal{E}_{\mathbf{z}}^{\epsilon}(f_{\lambda}) + \lambda \|f_{\lambda}\|_K^2] \leq 0$, we have

$$[\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}^{\epsilon})) + \lambda \|f_{\mathbf{z}}^{\epsilon}\|_K^2] - [\mathcal{E}_{\mathbf{z}}(f_{\lambda}) + \lambda \|f_{\lambda}\|_K^2] \leq q2^{q-1}\epsilon.$$

Then the desired conclusion holds. \square

In the above error decomposition, the first two terms S_1 and S_2 are called *sample error*. For the second term S_2 , we get the following estimation.

Corollary 2. *Assume that (2.5), there exists a subset $Z_{1,\delta}$ of Z^T with measure at least $1 - \frac{2\delta}{3}$ such that for any $\mathbf{z} \in Z_{1,\delta}$,*

$$S_2 \leq \frac{7q(1 + 5\|f_{\lambda}\|_{\infty}^q) \log \frac{3}{\delta}}{6T} + \frac{2^{q+1} \log \frac{3}{\delta}}{3T} + \left(\frac{2C_{\theta} \log \frac{3}{\delta}}{T} \right)^{\frac{1}{2-\theta}} + \mathcal{D}(\lambda). \quad (3.6)$$

Proof. we can decompose S_2 into two parts $S_2 = S_{2,1} + S_{2,2}$, where

$$\begin{aligned} S_{2,1} &= [\mathcal{E}_{\mathbf{z}}(f_{\lambda}) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\lambda}))] - [\mathcal{E}(f_{\lambda}) - \mathcal{E}(\pi(f_{\lambda}))], \\ S_{2,2} &= [\mathcal{E}_{\mathbf{z}}(\pi(f_{\lambda})) - \mathcal{E}_{\mathbf{z}}(f_q)] - [\mathcal{E}(\pi(f_{\lambda})) - \mathcal{E}(f_q)]. \end{aligned}$$

For $S_{2,1}$, we apply the one-side Bernstein inequality [2] to the random variable $\xi(z) = \psi_q(f_{\lambda}(x) - y) - \psi_q(\pi(f_{\lambda})(x) - y)$. For the continuity of the loss $\psi_q(u)$, it satisfies $0 \leq \xi \leq$

$q(\|f_\lambda\|_\infty + \|\pi(f_\lambda)\|_\infty + |y|)^{q-1} |\pi(f_\lambda)(x) - f_\lambda(x)| \leq q(2\|f_\lambda\|_\infty^{q-1} + 1)(1 + \|f_\lambda\|_\infty) \leq q(1 + 5\|f_\lambda\|_\infty^q)$. Noting that $|\xi - \mathbb{E}(\xi)| \leq q(1 + 5\|f_\lambda\|_\infty^q)$ and $\mathbb{E}(\xi - \mathbb{E}(\xi))^2 \leq q(1 + 5\|f_\lambda\|_\infty^q)\mathbb{E}(\xi)$, then there exists a subset $Z'_{1,\delta}$ of Z^T with measure at least $1 - \frac{\delta}{3}$ such that for any $\mathbf{z} \in Z'_{1,\delta}$,

$$S_{2,1} \leq \frac{7q(1 + 5\|f_\lambda\|_\infty^q) \log \frac{3}{\delta}}{6T} + \mathcal{E}(f_\lambda) - \mathcal{E}(\pi(f_\lambda)). \quad (3.7)$$

For $S_{2,2}$, we take the random variable $\xi(z) = \psi_q(\pi(f_\lambda)(x) - y) - \psi_q(f_q(x) - y)$ which is bounded by 2^q and estimate the variance by Lemma 1 with $f = \pi(f_\lambda)$. Applying the one-side Bernstein inequality again, we find that there exists a subset $\mathbf{z}''_{1,\delta}$ of Z^T with measure at least $1 - \frac{\delta}{3}$ such that for any $\mathbf{z} \in \mathbf{z}''_{1,\delta}$,

$$S_{2,2} \leq \frac{2^{q+1} \log \frac{3}{\delta}}{3T} + \left(\frac{2C_\theta \log \frac{3}{\delta}}{T} \right)^{\frac{1}{2-\theta}} + \mathcal{E}(\pi(f_\lambda)) - \mathcal{E}(f_q) \quad (3.8)$$

Combing the bound (3.7) and (3.8), we get the desired conclusion (3.6). \square

Denote $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. For $R \geq 1$, let $B_R = \{\mathbf{z} \in Z^T : \|f\|_K \leq R\}$.

Corollary 3. *Assume that (1.9) and (2.5). For any $f \in B_R$, there exists a subset $Z_{2,\delta}$ of Z^T with measure at least $1 - \frac{\delta}{3}$ such that for all $\mathbf{z} \in Z_{2,\delta}$,*

$$S_1 \leq \frac{1}{2} (\mathcal{E}(\pi(f)) - \mathcal{E}(f_q)) + 12\varepsilon^*(R, T, \delta/3)$$

where

$$\begin{aligned} \varepsilon^*(R, T, \delta/3) &\leq \left(2^{q+3} + (8C_\theta)^{\frac{1}{2-\theta}} \right) \log \frac{3}{\delta} T^{-\frac{1}{2-\theta}} \\ &\quad + \left(2^{q+3} C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)} + (8C_\theta C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)})^{\frac{1}{2+k-\theta}} \right) R^{\frac{k}{k+1}} T^{-\frac{1}{2+k-\theta}}. \end{aligned} \quad (3.9)$$

Proof. Consider the function set

$$\mathcal{G} = \{ \psi_q(\pi(f)(x) - y) - \psi_q(f_q(x) - y) : \|f\|_K \leq R \}.$$

A function from this set $g(z) = \psi_q(\pi(f)(x) - y) - \psi_q(f_q(x) - y)$ satisfies $\mathbb{E}g \geq 0$, $|g(z)| \leq 2^q$ and $\mathbb{E}g^2 \leq C_\theta (\mathbb{E}g)^\theta$ by (2.5). The continuity of the loss implies $|\psi_q(\pi(f)(x) - y) - \psi_q(f_q(x) - y)| \leq q(2 + \|f_q\|_\infty)^{q-1} |\pi(f)(x) - f_q(x)|$. Then

$$\mathcal{N}(\mathcal{G}, \varepsilon) \leq \mathcal{N} \left(B_1, \frac{\varepsilon}{q(2 + \|f_q\|_\infty)^{q-1} R} \right).$$

We apply the ratio probability inequality with the covering number in [16],

$$\begin{aligned} & \text{Prob}_{\mathbf{z} \in Z} \left\{ \sup_{\|f\|_K \leq R} \frac{[\mathcal{E}(\pi(f) - \mathcal{E}(f_q))] - [\mathcal{E}_{\mathbf{z}}(\pi(f) - \mathcal{E}_{\mathbf{z}}(f_q))]}{\sqrt{(\mathcal{E}(\pi(f) - \mathcal{E}(f_q))^\theta + \varepsilon^\theta)}} \leq 4\varepsilon^{1-\theta/2} \right\} \\ & \geq 1 - \mathcal{N} \left(B, \frac{\varepsilon}{q(2 + \|f_q\|_\infty)^{q-1} R} \right) \exp \left\{ -\frac{T\varepsilon^{2-\theta}}{2C_\theta + 2^{q+1}\varepsilon^{1-\theta}} \right\} \\ & \geq 1 - \exp \left\{ C_k \left(\frac{q(2 + \|f_q\|_\infty)^{q-1} R}{\varepsilon} \right)^k - \frac{T\varepsilon^{2-\theta}}{2C_\theta + 2^{q+1}\varepsilon^{1-\theta}} \right\}. \end{aligned}$$

We take $\varepsilon^*(R, T, \delta/3)$ to be the positive solution to the equation

$$C_k \left(\frac{q(2 + (\|f_q\|_\infty)R)}{\varepsilon} \right)^k - \frac{T\varepsilon^{2-\theta}}{2C_\theta + 2^{q+1}\varepsilon^{1-\theta/2}} = \log \frac{\delta}{3}.$$

It can be expressed as

$$\begin{aligned} & \varepsilon^{2+k-\theta} - \frac{2^{q+1}}{T} \log \frac{3}{\delta} \varepsilon^{1+k-\theta} - \frac{2C_\theta}{T} \log \frac{3}{\delta} \varepsilon^k - \frac{2^{q+1}C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)} R^k}{T} \varepsilon^{1-\theta} \\ & - \frac{2C_\theta C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)} R^k}{T} = 0. \end{aligned}$$

The positive solution $\varepsilon^*(R, T, \delta/3)$ to this equation can be bounded as

$$\begin{aligned} \varepsilon^*(R, T, \delta/3) & \leq \max \left\{ \frac{2^{q+3}}{T} \log \frac{3}{\delta}, \left(\frac{8C_\theta}{T} \log \frac{3}{\delta} \right)^{\frac{1}{2-\theta}}, \right. \\ & \left. \left(\frac{2^{q+3}C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)} R^k}{T} \right)^{\frac{1}{1+k}}, \left(\frac{8C_\theta C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)} R^k}{T} \right)^{\frac{1}{2+k-\theta}} \right\}. \end{aligned} \quad (3.10)$$

Then there exists a subset $Z_{2,\delta}$ of Z^T with measure at least $1 - \frac{\delta}{3}$ such that for all $\mathbf{z} \in Z_{2,\delta}$,

$$\sup_{\|f\|_K \leq R} \frac{[\mathcal{E}(\pi(f) - \mathcal{E}(f_q))] - [\mathcal{E}_{\mathbf{z}}(\pi(f) - \mathcal{E}_{\mathbf{z}}(f_q))]}{\sqrt{(\mathcal{E}(\pi(f) - \mathcal{E}(f_q))^\theta + (\varepsilon^*(R, T, \delta/3))^\theta)}} \leq 4(\varepsilon^*(R, T, \delta/3))^{1-\theta/2}.$$

For any $\mathbf{z} \in B(R) \cap Z_{2,\delta}$, we have

$$\begin{aligned} S_1 & \leq 4(\varepsilon^*(R, T, \delta/3))^{1-\theta/2} \sqrt{(\mathcal{E}(\pi(f) - \mathcal{E}(f_q))^\theta + (\varepsilon^*(R, T, \delta/3))^\theta)} \\ & \leq \frac{\theta}{2} (\mathcal{E}(\pi(f) - \mathcal{E}(f_q))) + (1 - \frac{\theta}{2}) 4^{1/(1-\theta/2)} \varepsilon^*(R, T, \delta/3) + 4\varepsilon^*(R, T, \delta/3) \\ & \leq \frac{1}{2} (\mathcal{E}(\pi(f) - \mathcal{E}(f_q))) + 12\varepsilon^*(R, T, \delta/3). \end{aligned}$$

Putting the above bounds into (3.10), then we get the desired conclusion (3.9). \square

4 Estimating Total Error by Iteration

This section is devoted to estimating total error $\|\pi(f_{\mathbf{z}}^\epsilon) - f_q\|_{L_{\rho_X}^r}$. To apply Corollary 2 and Corollary 3 for error analysis, we get the rough bound

$$\|f_{\mathbf{z}}^\epsilon\|_K \leq \lambda^{-\frac{1}{2}}, \quad \forall \mathbf{z} \in Z^T$$

by taking $f = 0$ in (1.4). This bound will be improved by iteration technique used in [14]. For $R > 0$, denote

$$\mathcal{W}(R) = \{\mathbf{z} \in Z^T : \|f_{\mathbf{z}}^\epsilon\|_K \leq R\}.$$

Lemma 3. *Take $\lambda = T^{-\alpha}$, $\epsilon = T^{-\eta}$ with $0 < \alpha \leq 1$, $0 < \eta \leq \infty$. Let $0 < \xi < 1$. If ρ satisfy the noise condition (1.8) and (1.6), (1.9) hold, then for any $0 < \delta < 1$, with confidence $1 - \delta$, there exists a subset V_R of Z^T with measure at most δ such that holds*

$$\|f_{\mathbf{z}}^\epsilon\|_K \leq 4A_2(1 + \sqrt{q2^q} + 2\sqrt{\mathcal{D}_0} + \sqrt{12q\mathcal{D}_0^{q/2}} + \sqrt{A_1})(\log \frac{3}{\xi})^2 \sqrt{\log \frac{3}{\delta}} T^\vartheta, \quad \forall \mathbf{z} \in \mathcal{W}(R) \setminus V(R), \quad (4.1)$$

where $\vartheta = \max \left\{ \frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)} + \xi, \frac{\alpha-\eta}{2}, \frac{\alpha(1-\beta)}{2}, \frac{\alpha}{2} + \frac{q(1-\beta)\alpha}{4} - \frac{1}{2}, \frac{\alpha}{2} - \frac{1}{2(2-\theta)} \right\}$.

Proof. Applying Corollary 2 and Corollary 3 with Lemma 2, we know that for any $\mathbf{z} \in \mathcal{W}(R) \cap Z_{1,\delta} \cap Z_{2,\delta}$, $R > 1$,

$$\begin{aligned} \mathcal{E}(\pi(f_{\mathbf{z}}^\epsilon)) - \mathcal{E}(f_q) + \lambda \|f_{\mathbf{z}}^\epsilon\|_K^2 &\leq q2^q\epsilon + 4\mathcal{D}(\lambda) + 12q\|f_\lambda\|_\infty^q T^{-1} \log \frac{3}{\delta} \\ &\quad + A_1 T^{-\frac{1}{2-\theta}} \log \frac{3}{\delta} + A_2 R^{\frac{k}{k+1}} T^{-\frac{1}{2+k-\theta}}. \end{aligned}$$

where A_1 and A_2 is given by

$$A_1 = 106 + 20C_\theta^{\frac{1}{2-\theta}}, A_2 = 2^{q+4}(C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)})^{\frac{1}{1+k}} + 16(C_\theta C_k q^k (2 + \|f_q\|_\infty)^{k(q-1)})^{\frac{1}{2+k-\theta}}.$$

Let V_R be a set whose measure is at most δ . Putting $\lambda = T^{-\alpha}$, $\epsilon = T^{-\eta}$ with $0 < \alpha \leq 1$, $0 < \eta \leq \infty$ and (1.6) into the above bound, then for any $R > 1$ we have

$$\|f_{\mathbf{z}}^\epsilon\|_K \leq a_T R^{\frac{k}{2+2k}} + b_T, \quad \mathbf{z} \in \mathcal{W}(R) \setminus V_R,$$

where the constants a_T and b_T are given by

$$a_T = \sqrt{A_2} T^{\frac{\alpha}{2} - \frac{1}{2(2+k-\theta)}}, b_T = \left\{ \sqrt{q2^q} + 2\sqrt{\mathcal{D}_0} + \sqrt{12q\mathcal{D}_0^{q/2} \log \frac{3}{\delta}} + \sqrt{A_1 \log \frac{3}{\delta}} \right\} T^\zeta$$

with $\zeta = \max \left\{ \frac{\alpha-\eta}{2}, \frac{\alpha(1-\beta)}{2}, \frac{\alpha}{2} + \frac{q(1-\beta)\alpha}{4} - \frac{1}{2}, \frac{\alpha}{2} - \frac{1}{2(2-\theta)} \right\}$. It follows that

$$\mathcal{W}(R) \subseteq \mathcal{W}(a_T R^{\frac{k}{2+2k}} + b_T) \cup V_R,$$

Let us apply the above relation iteratively to a sequence $\{R^{(j)}\}_{j=0}^J$ defined by $R^{(0)} = \lambda^{-\frac{1}{2}}$ and $R^{(j)} = a_T (R^{(j-1)})^{\frac{k}{2+2k}} + b_T$ where $J \in \mathbb{N}$ will be determined later. Then $\mathcal{W}(R^{(j-1)}) \subseteq \mathcal{W}(R^{(j)}) \cup V_{R^{(j-1)}}$. Noting that $\mathcal{W}(R^{(0)}) = Z^T$, then

$$Z^T = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}} \subseteq \cdots \mathcal{W}(R^{(J)}) \cup \left(\bigcup_{j=0}^{J-1} V_{R^{(j)}} \right).$$

As the measure of $V_{R^{(j)}}$ is at most δ , we know that the measure of $\bigcup_{j=0}^{J-1} V_{R^{(j)}}$ is at most $J\delta$. Hence $\mathcal{W}(R^{(J)})$ has measure at least $1 - J\delta$.

Denote $\Delta = \frac{k}{2+2k} \leq \frac{1}{2}$. The definition of the sequence $\{R^{(j)}\}_{j=0}^J$ implies that

$$R^{(J)} = a_T^{1+\Delta+\Delta^2+\cdots+\Delta^{J-1}} (R^{(0)})^{\Delta^J} + \sum_{j=1}^{J-1} a_T^{1+\Delta+\Delta^2+\cdots+\Delta^{j-1}} b_m^{\Delta^j} + b_m.$$

The first term

$$\begin{aligned} a_T^{1+\Delta+\Delta^2+\cdots+\Delta^{J-1}} (R^{(0)})^{\Delta^J} &= (A_2)^{\frac{1-\Delta^J}{2(1-\Delta)}} T^{\left(\frac{\alpha}{2} - \frac{1}{2(2+k-\theta)}\right) \frac{1-\Delta^J}{1-\Delta}} T^{\frac{\alpha}{2} \Delta^J} \\ &\leq A_2 T^{\frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)}} T^{\frac{1}{2+k-\theta} 2^{-J}}. \end{aligned}$$

Taking J be the smallest integer greater than or equal to $\log \frac{1}{\xi} / \log 2$. Then the upper bound

is estimated by $A_2 T^{\frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)} + \xi}$. The second term

$$\begin{aligned} \sum_{j=1}^{J-1} a_T^{1+\Delta+\Delta^2+\dots+\Delta^{j-1}} b_m^{\Delta^j} + b_m &\leq A_2 T^{\left(\frac{\alpha}{2} - \frac{1}{2(2+k-\theta)}\right) \frac{1-\Delta^j}{1-\Delta}} b_1^{\Delta^j} m^{\zeta \Delta^j} + b_1 m^\zeta \\ &\leq A_2 b_1 T^{\frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)}} \sum_{j=0}^{J-1} T^{\left(\zeta - \frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)}\right) \frac{k^j}{(2+2k)^j}}. \end{aligned}$$

where $b_1 = \sqrt{q2^q} + 2\sqrt{\mathcal{D}_0} + \sqrt{12q\mathcal{D}_0^{q/2} \log \frac{3}{\delta}} + \sqrt{A_1 \log \frac{3}{\delta}}$.

If $\zeta > \frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)}$, it is bounded by $A_2 b_1 J T^\zeta$. If $\zeta \leq \frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)}$, it is bounded by $A_2 b_1 J T^{\frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)}}$.

Thus we have

$$R^{(J)} \leq (A_2 + A_2 b_1 J) T^\vartheta,$$

where $\vartheta = \max\left\{\frac{[\alpha(2+k-\theta)-1](1+k)}{(2+k-\theta)(2+k)} + \xi, \zeta\right\}$. With confidence $1 - J\delta$, there holds

$$\|f_{\mathbf{z}}^\epsilon\|_K \leq R^{(J)} \leq A_2 (1 + \sqrt{q2^q} + 2\sqrt{\mathcal{D}_0} + \sqrt{12q\mathcal{D}_0^{q/2}} + \sqrt{A_1}) \sqrt{\log \frac{3}{\delta}} J T^\vartheta.$$

Noting $J \leq 2 \log \frac{3}{\xi}$, then we can get (4.1) by replacing δ by δ/J . \square

Now we can prove Theorem 2.

Proof of Theorem 2. By Lemma 3, there exists a subset $V_{R'} \subset Z^T$ with measure at most δ such that $Z^T \setminus V_{R'} \subseteq \mathcal{W}(R)$. Let R be the right side of (4.1). Applying Corollary 2 and Corollary 3 to R , then there exists another subset $V_R \subset Z^T$ with measure at most δ such that

$$\begin{aligned} \mathcal{E}(\pi(f)) - \mathcal{E}(f_q) &\leq q2^q \epsilon + 4\mathcal{D}(\lambda) + 12q \|f_\lambda\|_\infty^q T^{-1} \log \frac{3}{\delta} + A_1 T^{-\frac{1}{2-\theta}} \log \frac{3}{\delta} \\ &\quad + A_3 \left(\log \frac{3}{\xi}\right)^2 \sqrt{\log \frac{3}{\delta}} T^{\frac{k}{1+k} \vartheta - \frac{1}{2+k-\theta}}. \end{aligned}$$

where $A_3 = A_2(4A_2)^{\frac{k}{k+1}} (1 + \sqrt{q2^q} + 2\sqrt{\mathcal{D}_0} + \sqrt{12q\mathcal{D}_0^{q/2}} + \sqrt{A_1})$. By (2.1), we obtain that

$$\|\pi(f_{\mathbf{z}}^\epsilon) - f_q\|_{L_{\rho_X}^\epsilon} \leq C^* T^{-\Lambda}$$

where

$$C^* = C_r(q2^q + 4\mathcal{D}_0 + 12q\mathcal{D}_0^{q/2} + A_1 + A_3)^{\frac{1}{q+w}}$$

and Λ is given by (2). The restriction (1.12) ensures that $\Lambda > 0$. Replacing δ with $\delta/2$, we complete the proof of Theorem 2.

Now we are in the state of proving Theorem 1.

Proof of Theorem 1. We shall prove Theorem 1 by Theorem 2. First, we check the noise condition (1.8). Let the function $a(x) = \frac{1}{4}$ and $b(x) = 2^{2\varphi+1}$, $\forall x \in X$. For $s \in [0, a(x)] = [0, \frac{1}{4}]$, then

$$\rho_x(\{y : f_q(x) \leq y \leq f_q(x) + s\}) = \int_{f_q(x)}^{f_q(x)+s} \frac{d\rho_x(y)}{dy} dy = 2^{2\varphi+1} s^{\varphi+1}.$$

By similarity, $\forall s \in [0, \frac{1}{4}]$,

$$\rho_x(\{y : f_q(x) - s \leq y \leq f_q(x)\}) = 2^{2\varphi+1} s^{\varphi+1}.$$

So we say that ρ has a ∞ -average type $\varphi + 1$.

Since $f_q \in \mathcal{H}_K$ and $K \in C^\infty(X \times X)$, then (1.6) and (1.9) hold with $\beta = 1$ and $k = 0$. Thus, $\theta = \frac{2}{q+\varphi+1}$ and $r = q + \varphi + 1$. Noting that the choice of λ and ϵ satisfy (1.12) and $\Lambda > 0$. This complements our Theorem 1.

Proof of Corollary 1. It is an easy consequence of Theorem 2.

References

- [1] D. R. Chen, Q. Wu, Y. M. Ying and D. X. Zhou, Support vector machine soft margin classifiers: error analysis, *Journal of Machine Learning Research* **2** (2004) 1143–1175.
- [2] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1997.
- [3] T. Hu, J. Fan, Q. Wu and D. X. Zhou, Regularization schemes for minimum error entropy principle, *Analysis and Applications* **13**, 437, 2015, DOI: 10.1142/S0219530514500110.

- [4] P. J. Huber, *Robust Statistics*, Wiley, 1981.
- [5] T. Hu, D. H. Xiang and D. X. Zhou, Online learning for quantile regression and support vector regression, *Journal of Statistical Planning and Inference* **142** (2012), 3107–3122.
- [6] R. Koenker and G. Bassett, Regression quantiles, *Econometrica* **46** (1978), 33–50.
- [7] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.* **1** (2003), 17–41.
- [8] I. Steinwart, How to compare different loss functions and their risks, *Constr. Approx.* **26** (2007) 225–287.
- [9] I. Steinwart and A. Christmann, Estimating conditional quantiles with the help of the pinball loss, *Bernoulli* **17** (2011), 211–225.
- [10] I. Steinwart and A. Christmann, How support vector machines can estimate quantiles and the median, *Advances in Neural Information Processing Systems* **20** (2008), 305–312.
- [11] H. W. Sun and Q. Wu, Indefinite kernel network with dependent sampling, *Analysis and Applications* **11** (2013), DOI: 10.1142/S0219530513500206.
- [12] H. Z. Tong, D. R. Chen and L. Z. Peng, Analysis of support vector machines regression, *Found. Comput. Math.* **9** (2009), 243–257.
- [13] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [14] Q. Wu and D. X. Zhou, Analysis of support vector machine classification, *J. Comput. Anal. Appl.* **8** (2008), 99–119.
- [15] Q. Wu, Y. Ying and D. X. Zhou, Learning rates of least-square regularized regression, *Foundations of Computational Mathematics* **6** (2006) 171–192.
- [16] Q. Wu, Y. Ying and D. X. Zhou, Multi-kernel regularized classifiers, *J. Complexity*. **23** (2007), 108–134.

- [17] D. H. Xiang, T. Hu, and D. X. Zhou, Learning with varying insensitive loss, *Appl. Math. Letters* **24** (2011), 2107-2109.
- [18] D. H. Xiang, T. Hu, and D. X. Zhou, Approximation analysis of learning algorithms for support vector regression and quantile regression, *Journal of Applied Mathematics* *2012*, 2012.
- [19] D. H. Xiang and D. X. Zhou, Classification with gaussians and convex loss, *Journal of Machine Learning Research* **10** (2009), 1447–1468.
- [20] Y. Yao, *On Some Problems in the Mathematical Foundation of Learning*, M.Phil. Thesis, City University of Hong Kong, 2002.
- [21] T. Zhang, Covering number bounds of certain regularized linear function classes, *Journal of Machine Learning Research*, **2** (2002), 527–550.
- [22] D. X. Zhou, The covering number in learning theory, *J. Complexity* **18** (2002), 739–767.
- [23] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory*. **49** (2003), 1743–1752.